T-STAM:基于双流时空注意力机制的端到端的动作识别模型 *

石祥滨 [1,2]，李怡颖 [1†]，刘　芳 [2]，代　钦 [3]

(1. 辽宁大学 信息学院，沈阳 110036; 2. 沈阳航空航天大学 计算机学院，沈阳 110136; 3. 沈阳工程学院 信息学院，沈阳 110136)

摘　要：针对双流法进行视频动作识别时忽略特征通道间的相互联系、特征存在大量冗余的时空信息等问题，提出一种基于双流时空注意力机制的端到端的动作识别模型 T-STAM，实现了对视频关键时空信息的充分利用。首先，将通道注意力机制引入到双流基础网络中，通过对特征通道间的依赖关系进行建模来校准通道信息，提高特征的表达能力。其次，提出一种基于 CNN 的时间注意力模型，使用较少的参数学习每帧的注意力得分，重点关注运动幅度明显的帧。同时，提出一种多空间注意力模型，从不同角度计算每帧中各个位置的注意力得分，提取多个运动显著区域。接着，对时空特征进行融合进一步增强视频的特征表示。最后，将融合后的特征输入到分类网络，按不同权重融合两流输出得到动作识别结果。在数据集 HMDB51 和 UCF101 上的实验结果表明 T-STAM 能有效的识别视频中的动作。

关键词：动作识别；双流；通道信息；时空注意力；运动显著区域

中图分类号：TP391.41　　doi: 10.19734/j.issn.1001-3695.2020.02.0077

T-STAM: end-to-end action recognition model based on two-stream network with spatio-temporal attention mechanism

Shi Xiangbin[1, 2], Li Yiying[1†], Liu Fang[2], Dai Qin[3]

(1. *College of Information*, *Liaoning University*, *Shenyang* 110036, *China*; 2. *College of Computer Science*, *Shenyang Aerospace University*, *Shenyang* 110136, *China*; 3. *College of Information*, *Shenyang Institute of Engineering*, *Shenyang* 110136, *China*)

Abstract: Aiming at the problems that the action recognition methods based on two-stream ignore the inter-relationship between feature channels, and have large amount of redundant spatio-temporal information, this paper proposed an end-to-end action recognition model based on two-stream network with spatio-temporal attention mechanism (T-STAM) , which realized the full utilization of the key spatio-temporal information in the video. Firstly, this paper introduced the channel attention mechanism to the two-stream basic network, and calibrated the channel information by modeling the dependencies between feature channels to improve the ability of future expression. Secondly, this paper proposed a CNN-based temporal attention model to learn the attention score of each frame with fewer parameters, which can focuses on the frames with significant amplitude of motion. At the same time, it proposed a multi-spatial attention model, which calculated the attention score of each position in frame from different angles to extract motion saliency areas. Then, temporal and spatial features were fused to further enhance the feature representation of video. Finally, the fused features were input into the classification network, and the results of each stream are fused according to different weights to obtain the recognition results. The experimental results on the datasets HMDB51 and UCF101 show that T-STAM can effectively recognize actions in video.

Key words: action recognition; two-stream; channel information; spatio-temporal attention; motion saliency areas

0　引言

动作识别[1]在视频监控、智能家居、视频检索、人机智能交互等多种领域有广泛应用。视频具有环境复杂、视角和人体运动范围变换幅度较大等特点，这些特点使得对视频进行特征表示时，时间和空间上存在大量的冗余信息。因此，有效利用视频中运动幅度明显的帧上的关键区域(如物体、人与物体交互的身体部位等)的信息对动作识别至关重要。

视频中动作识别方法可以分为两类，传统方法[2,3]和基于深度学习的方法[4~9,121922~26]。传统方法在动作识别领域取得了一些进展，但其严重依赖于人工设计特征，算法的泛化能力不足。基于深度学习的方法能自动学习视频的特征进行分类。其中双流法[4~7]能有效结合视频中的时空信息，在性能上相对较优。Simonyan 等人[4]首次提出双流模型，将单帧图像和多帧密度光流场图像分别输入到空间流和时间流中，并对两流特征融合分类。Wang 等人[7]提出时态分段网络，使用稀疏采样和视频监督的策略，进一步的提升了识别精度。但双流法无法有效利用视频的关键时空信息。此外，它在提取视频特征时忽略了不同通道表示信息的差异性。为了获取视频中显著区域信息，文献[8~10]使用物体检测或者姿态估计提取视频中多个关键区域或者身体部位，再将其输入到网络中进行动作识别。但是预先对视频进行物体检测或姿态估计会增大整体计算代价，而且检测和估计的结果影响识别的性能。

基于注意力机制的动作识别方法[11~19]可以自动的学习视频中的关键信息。Hu 等人[11]设计了通道注意力网络，从通道上对特征进行建模来突出重点通道信息。Sharma 等人[12]提

收稿日期：2020-02-24；修回日期：2020-04-12　　基金项目：国家自然科学基金资助项目(61602320)

作者简介：石祥滨(1963-)，男，辽宁大连人，教授，硕导，博士，主要研究方向为分布式虚拟现实、网络游戏、数据库、模式识别、视频图像处理；李怡颖(1996-)，女(通信作者)，河南商丘人，硕士研究生，主要研究方向为图像处理、动作识别(1648646159@qq.com)；刘芳(1981-)，女，辽宁鞍山人，讲师，博士研究生，主要研究方向为计算机视觉、视频图像处理、动作识别；代钦(1981-)，男(蒙古族)，内蒙古兴安盟人，讲师，博士，主要研究方向为计算机视觉、视频图像处理、姿态估计.

出空间注意力模型突出每帧中的显著区域。Du 等人[14]采用 RNN 设计的时间注意力模型为不同的帧赋予相应的权重，可以有效利用视频的关键帧。Yang 等人[17]使用双向 LSTM 设计时空注意力模型。但是文献[12~19]有以下不足：a) 使用 RNN 或 LSTM 设计的时间注意力模型参数较多，且 RNN 具有固定的串行结构，必须按照时间的先后顺序来处理视频的帧，识别效率低。b) 在提取空间显著信息时，仅用一个空间注意力模型提取帧的多个运动区域，会产生提取区域信息不准确的问题。

针对上述问题，本文提出一种基于双流时空注意力机制的端到端的动作识别模型(end-to-end action recognition model based on two-stream network with spatio-temporal attention mechanism, T-STAM)。T-STAM 贡献如下：a) 将通道注意力融入到双流基础网络中，在兼顾双流特征的同时对特征的通道信息进行了重新校准，增强了特征的表达能力。b) 提出基于 CNN 的时间注意力模型来重点关注时域上判别力强的帧。与采用 RNN 的时间注意力模型相比，一方面，该模型基于 CNN 在视频的时间维度上计算每帧的注意力得分，模型的参数少且计算代价小；另一方面，采用 CNN 能实现多帧的并行运算，提高整体运行效率。c) 提出多空间注意力模型。采用多个模型从不同的角度学习每帧空间位置的权重，得到多个有判别性的运动区域，如人与物体交互，物体和身体运动部位等，减少了背景信息的干扰。将时空特征进行融合，进一步增强视频的特征表示。d) 在数据集 UCF101 和 HMDB51 上进行了实验验证。实验结果表明，T-STAM 是一种端到端的、高效的动作识别模型。

1　双流时空注意力机制的动作识别模型

视频可以看做由空间和时间两部分组成。空间上，RGB 图像包含场景和物体的外观信息，时间上，光流图像包括物体的运动信息。因此本文采用以 RGB 图像为输入的外观流和以光流图像为输入的运动流为基础进行设计。本文提出一种动作识别模型 T-STAM 来加强特征表示，能区分不同通道特征，并将注意力集中在判别力强的帧上的多个运动显著区域进行动作识别。T-STAM 整体结构如图 1 所示。为了获取合适的输入片段，T-STAM 对视频进行稀疏采样，具体实现为：将视频按等间隔分成 N 段，每段随机采样一帧，将采样帧的 RGB 图像和光流图像输入到双流网络中。

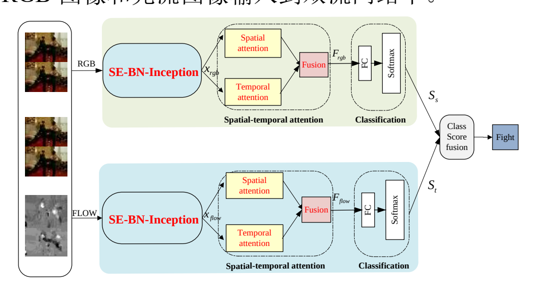


图 1　T-STAM 结构

Fig. 1　Structure of T-STAM

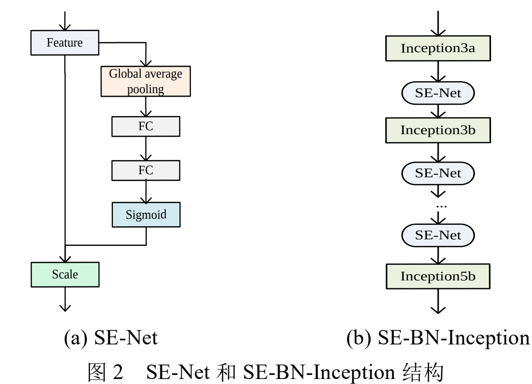
T-STAM 以外观流和运动流为基础，每流网络中均包含三个模块：SE-BN-Inception 模块、时空注意力模块和分类模块。SE-BN-Inception 模块能区分不同通道表示特征的差异性，从整体上提取到表达能力强的视频特征。经过本模块的外观流输出为 x_{rgb}，时间流输出为 x_{flow}；时空注意力模块能进一步加强视频的特征表示，通过时间注意力模型和多空间注意力模型分别在时间和空间上重点突出视频中识别力强的帧以及帧的多个运动显著区域。分类模块由一个 FC 层和一个 softmax 函数组成，将两流的时空特征 F_{rgb} 和 F_{flow} 分别输入到分类模块得到外观流输出 S_s 和运动流输出 S_t。按照不同权重融合两流的输出得到最终动作识别结果。

2　SE-BN-Inception 模块

使用卷积网络提取视频帧的特征时会产生多通道特征向量，向量的每个通道从特定方面描述当前帧，不同通道表示的信息重要程度并不相同。然而以往基于深度学习的方法提取特征时，忽略了不同通道表示信息的差异性，导致特征表示能力不强。而通道注意力机制能学习到每个特征通道的重要程度，按照重要程度提升对当前识别有用的通道特征，同时抑制识别力弱的通道特征。因此本文将通道注意力机制实现网络 SE-Net[11] (squeeze-and-excitation networks)引入到双流基础网络 BN-Inception[20]中得到 SE-BN-Inception 模块来校准不同通道信息，增强视频特征的表达能力。

SE-Net 如图 2(a)所示，网络具体实现如下：首先将输入特征沿着通道维度进行全局平均池化压缩特征。然后通过两个全连接层来建模通道间的依赖关系。第一个全连接层将输入通道维度降低为原来的 1/16 以减少计算量，之后通过 ReLU 激活函数增加非线性，第二个全连接层将通道回到原来的维度。再通过一个 sigmoid 函数获得归一化的权重，最后通过特征重定向操作将权重加权到每个通道的特征上。

SE-BN-Inception 模块结构如图 2(b)所示， BN-inception 包含 9 个 inception 操作，在每个 inception 后加入 SE-Net。由于全连接层的输出对空间和位置不够敏感，经过卷积层的输出在一定程度上保留了图像的空间结构，因此将 BN-inception 保留至最后一个卷积层。



(a) SE-Net　(b) SE-BN-Inception

图 2　SE-Net 和 SE-BN-Inception 结构

Fig. 2　Structure of SE-Net and SE-BN-Inception

3　时空注意力模块

时空注意力模块由基于 CNN 的时间注意力模型、多空间注意力模型以及时空特征的融合组成。时间注意力模型和多空间注意力模型分别从视频的时间和空间维度上重点关注关键帧和多个运动显著区域，时空特征的融合能有效结合提取的关键时空信息，进一步增强视频的特征表示，提高动作识别准确率。

3.1　基于 CNN 的时间注意力模型

动作是一个持续变化的过程，视频中不同的帧对识别动作的贡献程度并不相同。应该优先选择包含丰富信息，动作变化较明显的帧参加分类。时间注意力模型能为关键帧赋予更多的关注度。然而，以往的时间注意力模型[14~19]基于 RNN 设计实现，网络参数较多、结构复杂且无法随时间并行化。为了解决这个问题，本文提出一种基于 CNN 的时间注意力模型。采用 CNN 生成每帧的注意力得分，以注意力得分判断视频中每一帧相对于动作识别的重要性，选择性的关注重点帧，在时间维度上进一步的增强视频特征表示。本文设计的时间注意力模型不仅参数较少，结构简洁。还可以并行计算出所有帧的注意力得分，能充分利用 GPU 硬件的优势。基于 CNN 的时间注意力模型如图 3 所示。

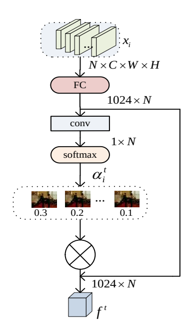


图 3　基于 CNN 的时间注意力模型

Fig. 3　Temporal attention model based on CNN

经过 SE-BE-Inception 模块后的特征为 $X=(x_1...x_i...x_N)$，$X\in R^{N\times C\times W\times H}$。$N$ 表示视频选用的帧数，C 代表特征维度，$W\times H$ 为特征图的网格单元数。对于视频第 i 帧的特征向量 x_i，先将其通过全连接层进行线性映射，映射后的特征为 $\hat{x}_i$，同一个视频帧的线性映射使用相同的参数，具体如式(1)。

$$\hat{x}_i=w_1x_i+b_1\quad i=1,2....N \tag{1}$$

其中 w_1、b_1 是模型中学习的参数，整个视频的映射特征为 $\hat{X}=(\hat{x}_1...\hat{x}_i...\hat{x}_N)$，$X\in R^{N\times D}$ ($D=256$)。将特征 $\hat{x}_i$ 通过一个卷积核大小为 1×1 的卷积层将视频特征维度变为 $1\times N$。沿视频帧的时间维度使用 softmax 函数得到视频的每一帧的时间注意力分数 α_i^t，计算如式(2)。

$$\alpha_i^t=\frac{\exp(conv(\hat{x}_i))}{\sum_{i=1}^{N}\exp(conv(\hat{x}_i))} \tag{2}$$

其中 $conv$ 代表卷积操作。α_i^t 表示第 i 帧对识别动作的贡献程度。获得第 i 帧的注意力得分 α_i^t 后，将其与特征 $\hat{x}_i$ 相乘得到第 i 帧的时间特征，对所有帧的时间特征求和得到整个视频的时间特征 f^t 如式(3)。

$$f^t=\sum_{i=1}^{N}\alpha_i^t x_i\qquad i=1,2....N \tag{3}$$

其中 $f^t\in R^{1\times D}$，它考虑到了视频中每个选取帧的重要程度。

3.2　多空间注意力模型

视频由序列图像组成，每帧空间上可分为运动显著区域以及其他区域。对于动作识别视频，运动显著区域通常是人体运动部位以及操作物体所在位置，如喝水这个动作，利用人的胳膊、头部区域以及杯子的特征就可以准确识别动作。因此，应重点关注每帧中的运动显著区域。以往采用物体检测[8,9]、姿态估计[10]等方法提取关键区域信息进行动作识别，工作量大且实现复杂。空间注意力机制[14~19]可以解决上述问题。然而，文献[14~19]仅使用一个空间注意力模型来提取不同显著区域信息，存在提取的部分显著区域不准确等问题。为了准确的提取帧的空间上与动作交互的不同区域信息，本文提出了多空间注意力模型，具体结构如图 4 所示。多空间注意力模型不是根据特征图的网格大小对输入图像进行空间上的分解，而是从多个角度来提取帧的空间信息，计算每帧中各个位置的注意力得分，进而找到不同的运动显著区域。这种学习方式可以减少背景等无关信息的干扰，缓解视频中人的姿态变化带来的问题，在空间上进一步增强视频的特征表示。空间注意力模型的个数代表着学习的运动显著区域数量，通过实验确定了空间注意力模型数的取值。

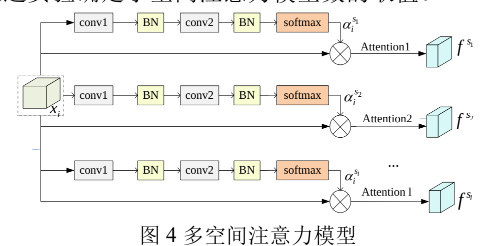


图 4 多空间注意力模型

Fig. 4　Multi-spatial attention model

本文采用多个空间注意力模型来提取帧的运动显著区域，每个模型主要由两个卷积层和一个 softmax 函数组成，对于第 j 个空间注意力模型，先将 X 经过一个 1×1 的卷积层和 tanh 激活函数把特征维度降至 $N\times F\times W\times H(F=256)$ 以减少计算代价。然后经过第二个卷积层得到特征 $c_i^{s_j}\ j\in(1,l)$，具体实现如式(4)。在每个卷积层后面都加入 BN(Batch Normalize)操作，引入 BN 操作可以解决协方差偏移问题，使训练更加稳定，BN 具体实现如式(5)。

$$c_i^{s_j}=BN(w_3(\tanh(BN(w_2x_i+b_2))))+b_3) \tag{4}$$

$$v^i=\frac{u^i-m}{\sqrt{\text{var}}}\times\alpha+\beta \tag{5}$$

式(4)中 w_2,w_3,b_2,b_3 是网络中可学习的参数。第二个卷积层的卷积核尺寸为 5×5，卷积步长为 1。$c_i^{s_j}\in R^{T\times l\times W\times H}$，$l$ 表示空间注意力模型个数。式(5)中 v^i 和 u^i 是输入和输出信号，α 和 β 是可训练参数，m 和 var 表示均值和方差。

将经过两个卷积层之后的特征 $c_i^{s_j}$ 输入到 softmax 函数计算第 i 帧中每个空间区域的概率得分 $\alpha_i^{s_{j,k}}(k\in R^{W\times H})$ 如式(6)。

$$\alpha_i^{s_{j,k}}=\frac{\exp c_i^{s_{j,k}}}{\sum_{k=1}^{W\times H}\exp c_i^{s_{j,k}}} \tag{6}$$

将 $\alpha_i^{s_{j,k}}$ 与每个映射特征进行元素相乘得到加权的空间特征。由于使用了 l 个空间注意力，每帧可提取 l 个空间特征。将每个视频选取帧的第 $j(j\in l)$ 个空间特征求和，得到整个视频的第 j 个空间特征 f^{s_j} 如式(7)。

$$f^{s_j}=\sum_{i=1}^{N}\sum_{k=1}^{W\times H}\alpha_i^{s_{j,k}}x_i^k \tag{7}$$

3.3　时空特征的融合

时空特征融合是结合视频提取的时间特征和空间特征来判断人的动作类别。融合的时空特征能表示关键帧的运动显著区域(人体运动部位、交互物体等)的变化信息，进一步增强特征的表达能力，对动作进行更准确的识别。比如打高尔夫球这个动作，通过时间注意力模型，挥球动作较明显的帧会获得更多的关注度。经过空间注意力模型提取到人的胳膊、高尔夫球杆、球等关键区域。结合时空特征可以重点关注挥球动作明显的帧上的多个运动显著区域，更好的识别动作。时空特征的融合如图 5 所示。

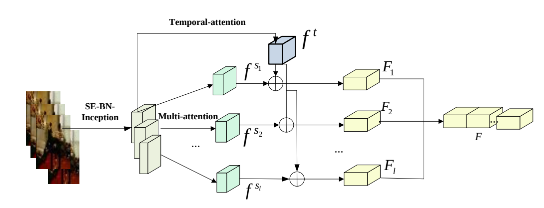


图 5　时空特征的融合

Fig. 5　Fusion of spatial-temporal features

每个视频分别获得 l 个空间特征 $f^{s_j}(j=1,2...l)$ 和一个时间特征 f^t，先将每个空间特征映射到时间特征上。即把视频的空间特征 f^{s_j} 分别和视频的时间特征 f^t 相加得到 l 个特征 F_l，然后将这 l 个特征连接起来得到视频的时空特征 F，具体如式(8)(9)所示。

$$F_l=f^{s_j}+f^t \tag{8}$$

$$F=concate(F_1,F_2...F_l) \tag{9}$$

其中：$concate$ 表示连接操作。

4　实验结果与分析

4.1　实验数据集和评价标准

本文采用的数据集为国际公开的两个基于视频的动作识别数据集：UCF101 和 HMDB51。

UCF101 数据集共包含 101 类动作、13320 个视频。该数

据集在动作的采集上具有较强的多样性，包括相机运动、物体外观运动、姿态变化和背景变化等。动作类别分为 5 组：人与物体交互、身体运动、人与人交互、演奏乐器和体育运动。该数据具有类内差异大、类间差异小等问题。HMDB51 数据集共有 6676 个视频、51 类动作。视频样本主要来源于电影、YouTube、Google 视频等公共数据，其中许多视频质量较差。因此，在这两个数据集上进行动作识别具有一定的挑战性。对于这两种数据集，本文采用官方划分的方式，即每个数据集都分成三个 split，每个 split 中 70%的视频是训练集，30%的视频是测试集。

本文采用 Top-1 识别准确率(以下简称识别准确率)作为评价标准。本文中每个数据集的识别准确率都是对其三个 split 的动作识别准确率进行加权平均求到的。

4.2 实验设置

本次实验在 GPU 版本的 pytorch 上执行。本文使用的 Backbone 是 BN-Inception，BN-Inception 模型是 GoogLeNet 模型的升级版，它在准确率和效率之间有着较好的平衡。本文使用在 ImageNet 数据集上预训练的模型参数对网络进行初始化。为了将光流数据和 RGB 数据保持一致，本文使用 Wang 等人[7]提供的工具提取光流。先采用 TV-L1 算法获取光流数据，然后通过线性变换将光流数据量化至[0,255]范围内。

训练阶段：先将输入帧的大小调整为 240×320，再采用固定角落裁剪和水平翻转，将裁剪区域的大小调整为 224×224。在分类网络的全连接层之前加入 dropout 层，外观流和运动流的 dropout 值分别设置为 0.8 和 0.7。通过小批量随机梯度下降算法优化参数，batchsize 为 32，权重衰减系数设置为 0.0005，动量设为 0.9。外观流的学习率初始值为 0.001，在 30 个 epoch 和 60 个 epoch 之后分别降低到原来的 1/10，共训练 80 个 epoch。运动流的学习率初始值为 0.001，在 190 个 epoch 和 300 个 epoch 之后分别降为原来的 1/10，共训练 340 个 epoch。

测试阶段：使用均值采样从每个样本中选取 25 帧图像，对于每帧图像，通过裁剪和翻转方式进行数据增强，获得 10 个测试样本，通过平均 10 个样本的输出类别概率得到分类结果。

4.3 实验分析

本节先对视频的不同分段数、不同空间注意力模型数、双流不同融合权重下的动作识别性能做了对比实验。然后，对加入通道注意力网络后的动作识别性能进行了实验分析。最后，将本文方法和 The-state-of-the-art 方法进行了比较，分析了本文方法的有效性。

4.3.1 不同视频分段数下的动作识别性能分析

本文使用 TSN 中的稀疏采样方法对视频中的帧进行采样，并将其作为网络的输入数据。为了分析不同视频分段数对动作识别性能的影响，本文在数据集 HMDB51 的第一个 split 上进行了对比实验。本文分别从视频中稀疏采样 3、4、5、6 个片段进行动作识别，在外观流上得到的实验结果如图 6 所示。实验结果表明，随着视频分段数的增加，识别准确率逐渐上升。当视频分段数为 6 时，网络的识别准确率最高。这是因为网络可以从不断增加的样本中学习到更多的信息。从图 6 可以看出，当视频分段数大于 5 时，随着分段数的增加，识别准确率的上升趋势逐渐变缓。而且由于电脑显存有限，无法测试更多的分段数目。因此本文将每个视频分成 6 段进行后续实验。

4.3.2 不同空间注意力模型数下的动作识别性能分析

本文提出的多空间注意力模型可以提取多个运动显著区域进行动作识别。随着空间注意力模型数的增加，提取的运动显著区域也逐渐增加。为了分析空间注意力模型数对动作识别性能的影响，本文在数据集 HMDB51 的第一个 split 上进行了对比实验，结果如图 7 所示。从图 7 可以看出，当空间注意力模型数小于 4 时，随着空间注意力模型数的增加，识别准确率逐渐提高。当空间注意力模型数为 4 时，动作识别性能最佳，当空间注意力模型数为 5 时，识别率下降。由于电脑显存有限，空间注意力模型数大于 5 时实验无法运行。因此本文采用 4 个空间注意力模型进行后续实验。

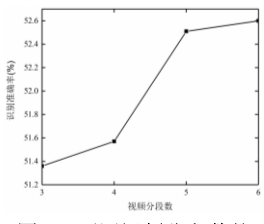


图 6　不同视频分段数的动作识别准确率比较

Fig. 6　Comparison of recognition accuarcy of different video segements

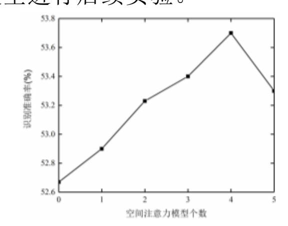


图 7　不同空间注意力模型数的识别准确率比较

Fig. 7　Comparison of recognition accuarcy of different spatial attention numbers

4.3.3 双流网络不同融合权重下的动作识别性能分析

本文通过实验分析外观流和运动流的不同融合权重对动作识别性能的影响，结果如表 1 所示。从表 1 可以看出，仅采用运动流比外观流识别准确率高，双流融合比单流的效果好。当外观流和运动流按 1/4 和 3/4 的权重进行融合时，动作识别结果最好。因此本文选取外观流和运动流融合权重为 1:3 进行后续实验。

表 1　双流不同融合权重的识别准确率对比

Tab. 1　Comparison of recognition accuracy of different fusion weights in two streams　/%

融合方法	识别率	融合方法	识别率
仅 RGB 流	53.33	1/2RGB 流和 1/2 光流	67.53
仅光流	64.93	1/4RGB 流和 3/4 光流	71.8
1/3RGB 流和 2/3 光流	69.38		

4.3.4 加入通道注意力网络后的动作识别性能分析

为了验证通道注意力网络的有效性，将加入 SE-Net 的 TSN[7]的模型与 TSN 模型在两个数据集上的识别准确率进行对比，融入后的模型采用与 TSN 相同的实验参数。对比结果如表 2 所示，可以看出，与 TSN 相比，融入 SE-Net 的 TSN 模型在数据集 UCF101 和 HMDB51 的识别准确率分别有 0.2%和 1.3%的提升。说明融入通道注意力网络能突出视频中有区分性的通道信息，增强特征的表达能力，提高动作识别的性能。

表 2　融入 SE-Net 的 TSN 模型与 TSN 识别准确率比较

Tab. 2　Comparison of recognition accuracy between TSN integrated with SE-Net and TSN　/%

方法	Backbone architecture	UCF101	HMDB51
TSN[7]	BN-Inception	94.9	69.4
SE-Net+TSN	SE-BN-Inception	95.1	70.7

4.3.5 与 the-state-of-the-art 方法的对比实验

1) 与使用注意力的动作识别方法的对比实验分析

为了验证本文提出的时空注意力模型的有效性，将本文算法 T-STAM(不含 SE-Net)与其他使用注意力机制的动作识别方法进行比较，实验结果如表 3 所示。通过表 3 可以看出，本文的方法具有更高的准确率。a)与采用 RNN 的方法生成的时间注意力模型 Temporal attention[13]相比，T-STAM(不含 SE-Net)在数据集 HMDB51 上准确率有了 6.3%的提升。这是因为 Temporal attention[13]仅仅提取了关键帧，而本文既提取了关键帧，还关注了空间维度上的运动显著区域，说明时空信息的结合能有效地提高识别精度。b)T-STAM(不含 SE-Net)的识别效果优于以 BN-Inception 为 Backbone 的时空注意力模型 RSTAN[14]和 ISTPAN[15]，说明在使用相同的 Backbone 下，

本文提出的时空注意力模型虽然结构简单，但能更有效地提取到视频的关键时空信息。c)与 Attention cluster[16]、使用双向 LSTM 设计的注意力网络 Bi-LSTM attention[17]、基于残差的时空注意力的模型 R-STAN[18]相比，T-STAM(不含 SE-Net)具有更好的性能。文献[16~18]都以 ResNet 为 Backbone 进行动作识别，ResNet 的网络性能优于 BN-Inception，但本文采用 BN-Inception 为 Backbone 的仍然获得了较好的识别效果。这说明本文提出的时空注意力模型可以有效弥补 BN-Inception 的不足，准确提取视频中的关键时空信息，提高动作识别的准确率。d)加入 SE-Net 之后，本文的 T-STAM 在两个数据集上的识别准确率有了进一步的提升，这说明结合通道注意力网络后，T-STAM 可以通过校准特征通道的信息来提高动作识别的性能。

表 3　与使用注意力的动作识别方法的识别准确率比较

Tab. 3　Comparison of recognition accuracy with attention_based action recognition methods　/%

方法	Backbone architecture	UCF101	HMDB51
Temporal attention[13]	BN-Inception	93.3	65.0
RSTAN[14]	BN-Inception	94.6	70.5
ISTPAN[15]	BN-Inception	94.8	69.6
Attention cluster[16]	ResNet-152	94.6	69.2
Bi-LSTM attention[17]	ResNet-152	94.8	71.9
R-STAN[18]	ResNet-152	94.5	68.7
本文(without SE-Net)	BN-Inception	95.3	71.3
本文(T-STAM)	SE-BN-Inception	95.7	71.9

2) 与近年来经典的动作识别方法的对比实验分析

为了进一步的验证本文方法，将 T-STAM 与一些经典的动作识别方法进行比较，结果如表 4 所示。从表 4 可以看出：a)与传统方法 IDT[2]相比，T-STAM 的识别准确率较高。这说明本文提出的时空注意力模型能有效地提取出视频中关键的时空信息，提高动作识别的效果。且 T-STAM 采用的端到端的结构使得计算更加简洁。b)与双流模型[5]、时态分段网络 TSN[7]相比，T-STAM 在数据集 UCF101 识别准确率分别提升了 3.2%和 0.8%，在数据集 HMDB51 上识别准确率分别提升了 6.5%和 2.5%。这说明 T-STAM 在双流的基础上加入的时空注意力模型，可以有效地提取到关键帧上更多的运动特征，通过这些信息能更准确的识别视频中的动作。c)与传统特征提取和卷积网络相结合的分类算法 TDD[21]、训练较深的 C3D 网络[22]、时空残差模型 ST-ResNet[23]、时空金字塔模型[24]以及 ARTNet[25]、TSM[26]等比较结果可以看出，T-STAM 的识别效果更优。这说明 T-STAM 兼顾了双流特征，对通道特征进行的重新校准突出了重点通道信息，提出的时空注意力模型充分的挖掘了视频的关键时空信息，获取了表达能力增强的视频特征，建立了全面的动作描述。

表 4　与近年经典的动作识别方法的识别准确率比较

Tab. 4　Comparison of recognition accuracy with classical action recognition methods in recent years　/%

方法	UCF101	HMDB51
IDT[2]	85.9	57.2
Two-stream fusion[5]	92.5	65.4
TSN[7]	94.9	69.4
TDD[21]	90.3	63.2
C3D[22]	82.3	56.8
ST-ResNet[23]	93.4	66.4
ST-pyramid[24]	94.6	68.9
ARTNet[25]	94.3	70.9
TSM[26]	94.5	70.7
本文(T-STAM)	95.7	71.9

5　结束语

针对双流法忽略特征不同通道信息的差异性、无法区分视频的冗余帧、背景等时空信息，造成整体特征表达能力不强，识别率不高的问题，本文提出一种基于双流时空注意力机制的端到端的动作识别模型。本文先将通道注意力融入双流结构，通过对通道特征的建模来校准通道信息，提高视频中特征表达能力。再设计基于 CNN 的时间注意力模型和多空间注意力模型来重点关注判别力强的帧上的多个运动显著区域，进一步增强视频的特征表示。本文在数据集 UCF101 和 HMDB51 上进行了对比实验，与近年来的先进方法相比，本文取得了较高的识别准确率。实验结果说明本文模型能有效的区分不同通道特征，将注意力集中在视频的关键时空信息上，更准确的识别视频中的动作。但本文的运动流和外观流采用了相同的网络结构，而人类对运动信息和外观信息的理解是两个不同的过程，因此两流的网络结构应该有所区分。今后的工作中将会在运动流和外观流设计不同的网络结构进行研究，也会探索与其他深度学习模型相结合，进一步提高识别准确率。

参考文献：

[1] 李瑞峰，王亮亮，王珂．人体动作行为识别研究综述 [J]．模式识别与人工智能，2014, 27 (1): 35-48. (Li Ruifeng, Wang Liangliang, Wang Ke. A survey of human body action recognition [J]. Pattern Recognition and Artificial Intelligence, 2014, 27 (1): 35-48.)

[2] Wang Heng, Schmid C. Action recognition with improved trajectories [C]// Proc of IEEE International Conference on Computer Vision. 2013: 3551-3558.

[3] 张杰，吴剑章，汤嘉立，等．基于时空图像分割和交互区域检测的人体动作识别方法 [J]．计算机应用研究，2017, 34(1): 302-305. (Zhang Jie, Wu Jianzhang, Tang Jiali, *et al.* Human action recognition method based on spatio-temporal image segmentation and interactive area detection [J]. Application Research of Computers, 2017, 34(1): 302-305.)

[4] Simonyan K, Zisserman A. Two-stream convolutional networks for action recognition in videos [J]. Advances in Neural Information Processing Systems. 2014, 1 (4): 568-576.

[5] Feichtenhofer C, Pinz A, Zisserman A. Convolutional two-stream network fusion for video action recognition [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2016: 1933-1941.

[6] Yue-Hei Ng J, Hausknecht M, Vijayanarasimhan S, *et al.* Beyond short snippets: Deep networks for video classification [C]// Pro of IEEE Conference on Computer Vision and Pattern Recognition. 2015: 4694-4702.

[7] Wang Limin, Xiong Yuanjun, Wang Zhe, *et al.* Temporal segment networks: Towards good practices for deep action recognition [C]// Proc of European Conference on Computer Vision. Cham: Springer, 2016: 20-36.

[8] Wang Yifan, Song Jie, Wang Limin, *et al.* Two-Stream SR-CNNs for Action Recognition in Videos [C]// BMVC. 2016.

[9] Tu Zhigang, Xie Wei, Qin Qianqing, *et al.* Multi-stream CNN: Learning representations based on human-related regions for action recognition [J]. Pattern Recognition, 2018, 79: 32-43.

[10] Chéron G, Laptev I, Schmid C. P-cnn: Pose-based cnn features for action recognition [C]// Proc of IEEE International Conference on Computer Vision. Piscataway, NJ: IEEE Press, 2015: 3218-3226.

[11] Hu Jie, Li Shen, Gang Sun. Squeeze-and-excitation networks [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. 2018:

7132-7141.

[12] Sharma S, Kiros R, Salakhutdinov R. Action recognition using visual attention [C]// Proc of International Conference on Machine Learning. 2015: 48-57

[13] Liu Zhikang, Tian Ye, Wang Zilei. Improving human action recognition by temporal attention [C]// Proc of IEEE International Conference on Image Processing. 9, 2017: 870-874.

[14] Du Wenbin, Wang Yali, Yu Qiao. Recurrent spatial-temporal attention network for action recognition in videos [J]. IEEE Transactions on Image Processing, 2017, 27 (3): 1347-1360.

[15] Du Yang, Yuan Chunfeng, Li Bing, *et al.* Interaction-aware spatio-temporal pyramid attention networks for action classification [C]// Proc of European Conference on Computer Vision. Cham: Springer, 2018: 373-389.

[16] Long Xiang, Gan Chuang, De Melo G, *et al.* Attention clusters: Purely attention based local feature integration for video classification [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2018: 7834-7843.

[17] Yang Haodong, Zhang Jun, Li Shuohao, *et al.* Bi-direction hierarchical LSTM with spatial-temporal attention for action recognition [J]. Journal of Intelligent & Fuzzy Systems, 2019, 36 (1): 775-786.

[18] Liu Quanle, Che Xiangjiu, Mei Bie. R-STAN: Residual spatial-temporal attention network for action recognition [J]. IEEE Access, 2019, 7: 82246-82255.

[19] Yan Shiyang, Smith J S, Lu Wenjin, *et al.* CHAM: action recognition using convolutional hierarchical attention model [C]// Proc of IEEE International Conference on Image Processing. 2017: 3958-3962.

[20] Ioffe S, Szegedy C. Batch normalization: Accelerating deep network training by reducing internal covariate shift [J]. Computer Science. 2015: 448-456.

[21] Wang Limin, Yu Qiao, Tang Xiaoou. Action recognition with trajectory-pooled deep-convolutional descriptors [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2015: 4305-4314.

[22] Tran D, Bourdev L, Fergus R, *et al.* Learning spatiotemporal features with 3D convolutional networks [C]// Proc of IEEE International Conference on Computer Vision. Washington DC: IEEE Computer Society, 2015: 4489-4497

[23] Feichtenhofer C, Pinz A, Wildes R. Spatiotemporal residual networks for video action recognition [C]// Advances in Neural Information Processing Systems. 2016: 3468-3476.

[24] Wang Yunbo, Long Mingsheng, Wang Jianmin, *et al.* Spatiotemporal Pyramid Network for Video Action Recognition [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition Piscataway, NJ: IEEE Press, 2017.

[25] Wang Limin, Li Wei, Li Wen, *et al.* Appearance-and-relation networks for video classification [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2018: 1430-1439.

[26] Lin Ji, Gan Chuang, Han Song. Tsm: Temporal shift module for efficient video understanding [C]// Proc of IEEE International Conference on Computer Vision. 2019: 7083-7093.